CLAIMS

What is claimed is:

1. A method of identifying a relationship between one or more candidate biomolecules and one or more reference biomolecules, the method comprising:

(a) inputting to a computer a query set describing the one or more candidate biomolecules;

(b) comparing the query set with a target database describing the one or more reference biomolecules, wherein the one or more reference biomolecules are grouped into one or more buckets, and wherein the one or more reference biomolecules of each bucket share a common property;

(c) counting a number of matches between each query set and each bucket of the target database; and

(d) statistically analyzing each match, wherein the presence of a statistically significant match identifies a relationship between the query set and a bucket of the target database.

2. The method of claim 1, wherein the query set comprises one or more sequences.

3. The method of claim 2, wherein'the one or more sequences are selected from the group consisting of a DNA sequence, an RNA sequence, and a protein sequence.

4. The method of claim 2, wherein the one or more sequences are extracted from one genetic region.

5. The method of claim 1, wherein the one or more candidate biomolecules and the one or more reference biomolecules are all selected from the group consisting of proteins, nucleic acids, and small molecules.

6. The method of claim 1, wherein the comparing comprises employing an equivalence algorithm based on identity of name, accession, or other identifier associated with biomolecule.

7. The method of claim 1, wherein the comparing comprises employing a BLAST-based algorithm to identify similarities or identities in two or more sequences.

8. The method of claim 1, wherein the counting comprises applying one or more principles chosen from the group consisting of:

(a) each query set candidate sequence can match at most one reference sequence in any given bucket;

(b) each query set candidate sequence can possess a match in one or more different buckets; and

(c) once a candidate sequence in the query set matches a specific bucket reference sequence in the target database, any subsequent matches of that same candidate sequence to other reference sequences in that bucket do not increase the match count for the bucket.

9. The method of claim 1, wherein the statistically analyzing comprises computing one or more statistics for each bucket.

10. The method of claim 8, further comprising sorting the one or more statistics by increasing or decreasing significance.

11. The method of claim 1, further comprising outputting a webpage with results of the statistical analysis, the webpage comprising one or more hyperlinks.

12. A computer-readable storage device embodying a program of instructions executable by a computer to perform method steps for identifying

a relationship between one or more candidate biomolecules and one or more reference biomolecules, the method steps comprising:

(a) inputting to a computer a query set describing one or more candidate biomolecules;

(b) comparing the query set with a target database describing one or more reference biomolecules, the one or more reference biomolecules of the target database grouped into one or more buckets, wherein the one or more reference biomolecules of each bucket share a common property;

(c) counting a number of matches between each query set and each bucket of the target database; and

(d) statistically analyzing each match, wherein the presence of a statistically significant match identifies a relationship between a query set and one or more buckets of a target database.

13. The computer-readable storage device of claim 12, wherein the query set comprises one or more candidate sequences.

14. The computer-readable storage device of claim 13, wherein the one or more candidate sequences are selected from the group consisting of a DNA sequence, an RNA sequence, and a protein sequence.

15. The computer-readable storage device of claim 13, wherein the one or more candidate sequences are extracted from one genetic region.

16. The computer-readable storage device of claim 12, wherein the one or more candidate biomolecules and the one or more reference biomolecules are all selected from the group consisting of proteins, nucleic acids, and small molecules.

17. The computer-readable storage device of claim 12, wherein the comparing comprises employing a BLAST-based algorithm to identify similarities or identities in two or more sequences.

18. The computer-readable storage device of claim 12, wherein the comparing comprises employing a equivalence algorithm based on identity of name, accession, or other identifier associated with biomolecule.

19. The computer-readable storage device of claim 12, wherein the counting comprises applying one or more principles chosen from the group consisting of:

(a) each query set candidate sequence can match at most one reference sequence in any given bucket;

(b) each query set candidate sequence can possess a match in one or more different buckets; and

(c) once a candidate sequence in the query set matches a specific bucket reference sequence in the target database, any subsequent matches of that same candidate sequence to other reference sequences in that bucket do not increase the match count for the bucket.

20. The computer-readable storage device of claim 12, wherein the statistically analyzing comprises computing one or more statistics for each match.

21. The computer-readable storage device of claim 20, further comprising sorting the one or more statistics by increasing or decreasing significance.

22. The computer-readable storage device of claim 12, further comprising outputting a webpage with results of the statistically analyzing, the webpage comprising one or more hyperlinks.

23. A method of identifying a relationship between two or more region sets, each region set describing one or more candidate biomolecules, and a target database describing one or more reference biomolecules grouped into one or more buckets, the method comprising:

(a) providing a query set describing two or more region sets, each region set comprising one or more candidate biomolecule sequences extracted from one region;

(b) comparing the query set with target database sequences describing one or more reference biomolecule sequences, wherein the target database sequences are grouped into one or more buckets, and wherein the one or more reference biomolecules of each bucket share a common property;

(c) counting a number of matches between each query set and each bucket of the target database; and

(d) statistically analyzing each match, wherein the presence of a statistically significant match identifies a relationship between the query set and the bucket of the target database.

24. The method of claim 23, wherein the one or more biomolecule sequences are selected from the group consisting of protein sequences and nucleic acid sequences.

25. The method of claim 24, wherein the nucleic acid sequences are selected from the group consisting of a DNA sequence and an RNA sequence.

26. The method of claim 23, wherein the comparing comprises employing a equivalence algorithm based on identity of name, accession, or other identifier associated with biomolecule.

27. The method of claim 23, wherein the comparing comprises employing a BLAST-based algorithm to identify similarities or identities in two or more sequences.

28. The method of claim 23, wherein the counting comprises applying one or more principles chosen from the group consisting of:

(a) each query set candidate sequence can match at most one reference sequence in any given bucket;

(b) each query set candidate sequence can possess a match in one or more different buckets; and

(c) once a candidate sequence in the query set matches a specific bucket reference sequence in the target database, any subsequent matches of that same candidate sequence to other reference sequences in that bucket do not increase the match count for the bucket.

29. The method of claim 23, wherein the statistically analyzing comprises computing one or more statistics for each match.

30. The method of claim 29, further comprising sorting the one or more statistics by increasing or decreasing significance.

31. The method of claim 30, further comprising further comprising outputting a webpage with results of the statistically analyzing, the webpage comprising one or more hyperlinks.

32. The method of claim 23, the method further comprising:

(a) constructing a plurality of replicates of the one or more query sets;

(b) modeling the replicates at random chromosomal locations to form a random location data set;

(c) processing the random location data set by following the steps of claim 23;

(d) quantifying the number of times each match is found to surpass a predetermined threshold to form a statistically significant set of random location matches; and

(e) comparing the statistically significant set of random location matches to the statistically significant relationship of claim 23.

33. A computer-readable storage device embodying a program of instructions executable by a computer to perform method steps for identifying a relationship between two or more region sets, each region set each region set describing one or more candidate biomolecules, and a target database describing one or more reference biomolecules grouped into one or more buckets, the method steps comprising:

(a) providing a query set describing two or more region sets, each region set comprising one or more candidate biomolecule sequences extracted from one genetic region;

(b) comparing the query set with target database sequences describing one or more reference biomolecule sequences, wherein the target database sequences grouped into one or more buckets, and wherein the one or more reference biomolecules of each bucket share a common property;

(c) counting a number of matches between each query set and each bucket of the target database; and

(d) statistically analyzing each match, wherein the presence of a statistically significant match identifies a relationship between the query set and the bucket of the target database.

34. The computer-readable storage device of claim 33, wherein the one or more candidate biomolecule sequences and the one or more reference biomolecules sequences are all selected from the group consisting of protein sequences and nucleic acid sequences.

35. The computer-readable storage device of claim 34, wherein the nucleic acid sequences are selected from the group consisting of a DNA sequence and an RNA sequence.

36. The computer-readable storage device of claim 33, wherein the comparing comprises employing a BLAST-based algorithm to identify similarities or identities in two or more sequences.

37. The computer-readable storage device of claim 33, wherein the counting comprises applying one or more principles chosen from the group consisting of:

(a) each query set candidate sequence can match at most one reference sequence in any given bucket;

(b) each query set candidate sequence can possess a match in one or more different buckets; and

(c) once a candidate sequence in the query set matches a specific bucket reference sequence in the target database, any subsequent matches of that same candidate sequence to other reference sequences in that bucket do not increase the match count for the bucket.

38. The computer-readable storage device of claim 33, wherein the statistically analyzing comprises computing one or more statistics for each match.

39. The computer-readable storage device of claim 38, further comprising sorting the one or more statistics by increasing or decreasing significance.

40. The computer-readable storage device of claim 39, further comprising outputting a webpage with results of the statistically analyzing, the webpage comprising one or more hyperlinks.

41. The computer-readable storage device of claim 33, the method steps further comprising:

(a) constructing a plurality of replicates of the one or more query sets;

(b) modeling the replicates at random chromosomal locations to form a random location data set;

(c) processing the random location data set by following the steps of claim 33;

(d) quantifying the number of times each match is found to surpass a predetermined threshold to form a statistically significant set of random location matches; and

(e) comparing the statistically significant set of random location matches to the statistically significant relationship of claim 33.

42. A computer-readable medium having stored thereon a data structure having multiple data fields comprising:

(a) a first data field containing data representing a bucket;

(b) a second data field containing data representing a name for the bucket; and

(c) a third data field containing data representing a list of members of the bucket, wherein the members have a common property.

43. The computer-readable medium of claim 42, further comprising a data field containing data representing an organism from which each of the members of the bucket are derived.

44. The computer-readable medium of claim 42, further comprising a data field containing data representing a bucket source.

45. The computer-readable medium of claim 44, further comprising a data field containing data representing data for creating the bucket.

46. The computer-readable medium of claim 44, further comprising a Perl script that parses data and creates a bucket file.

47. The computer-readable medium of claim 42, further comprising a data field containing data representing standard nomenclature for each reference biomolecule that is a member of the bucket.

48. The computer-readable medium of claim 42, further comprising a data field containing data representing a sequence for a member of the bucket.

49. The computer-readable medium of claim 48, wherein the data representing a sequence for a member of the bucket is a nucleic acid sequence.

50. The computer-readable medium of claim 48, wherein the data representing a sequence for a member of the bucket is an amino acid sequence.

51. The computer-readable medium of claim 48, wherein:

(a) the data representing a sequence for a member of the bucket is an identification number, and

(b) the identification number allows for retrieval of the sequence.

52. The computer-readable medium of claim 51, wherein the identification number is an accession number wherein the accession number allows for retrieval of the sequence from a database.

53. The computer-readable medium of claim 52, wherein the database is chosen from the group consisting of a publicly available database and a private database.

54. The computer-readable medium of claim 53, wherein the publicly available database is chosen from the group consisting of SwissProt, TrEMBL, and NCBI.

55. A method of making a target database, the method comprising:

(a) identifying a source of informative content;

(b) arranging informative content from the source of informative content into a set of buckets, wherein the buckets are given names;

(c) gathering the names of the buckets and a list of biomolecules present in each bucket; and

(d) creating and loading into a database data fields containing data representing:

(i) the set of buckets;

(ii) the list of biomolecules present in each bucket; and

(iii) a description for each biomolecule present in each bucket.

56. The method of claim 55, wherein the source of informative content is a publicly available database.

57. The method of claim 56, wherein the publicly available database is chosen from the group consisting of SwissProt, TrEMBL, and NCBI.

58. The method of claim 55, wherein the gathering is accomplished using a source-specific parsing script.

59. The method of claim 55, wherein the creating and loading is accomplished using a database loading script.

60. The method of claim 55, wherein the data representing a description for each biomolecule present in each bucket is selected from the group consisting of a nucleic acid sequence, an amino acid sequence, or an identification number, wherein the identification number allows for retrieval of a nucleic acid sequence or an amino acid sequence.

61. A computer-readable storage device embodying a program of instructions executable by a computer to perform method steps for making a target database, the method steps comprising:

(a) identifying a source of informative content;

(b) arranging informative content from the source of informative content into a set of buckets, wherein the buckets are given names;

(c) gathering the names of the buckets and a list of biomolecules present in each bucket; and

(d) creating and loading into a database data fields containing data representing:

(i) the set of buckets;

(ii) the list of biomolecules present in each bucket; and

(iii) a description for each biomolecule present in each bucket.

62. The computer-readable storage device of claim 61, wherein the source of informative content is a publicly available database.

63. The computer-readable storage device of claim 62, wherein the publicly available database is chosen from the group consisting of SwissProt, TrEMBL, and NCBI.

64. The computer-readable storage device of claim 61, wherein the gathering is accomplished using a source-specific parsing script.

65. The computer-readable storage device of claim 61, wherein the creating and loading is accomplished using a database loading script.

66. The computer-readable storage device of claim 61, wherein the data representing a description for each biomolecule present in each bucket is selected from the group consisting of a nucleic acid sequence, an amino acid sequence, or an identification number, wherein the identification number allows for retrieval of a nucleic acid sequence or an amino acid sequence.